PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
H04Q 11/04, H04L 29/06, H04Q 3/00, G06F 11/14

A1

(11) International Publication Number: **WO 99/27750**

(43) International Publication Date: 3 June 1999 (03.06.99)

(21) International Application Number: PCT/CA98/01074

(22) International Filing Date: 19 November 1998 (19.11.98)

(30) Priority Data:
2,221,661 20 November 1997 (20.11.97) CA

(71) Applicant *(for all designated States except US)*: CROSSKEYS SYSTEMS CORPORATION [CA/CA]; 350 Terry Fox Drive, Kanata, Ontario K2K 2W5 (CA).

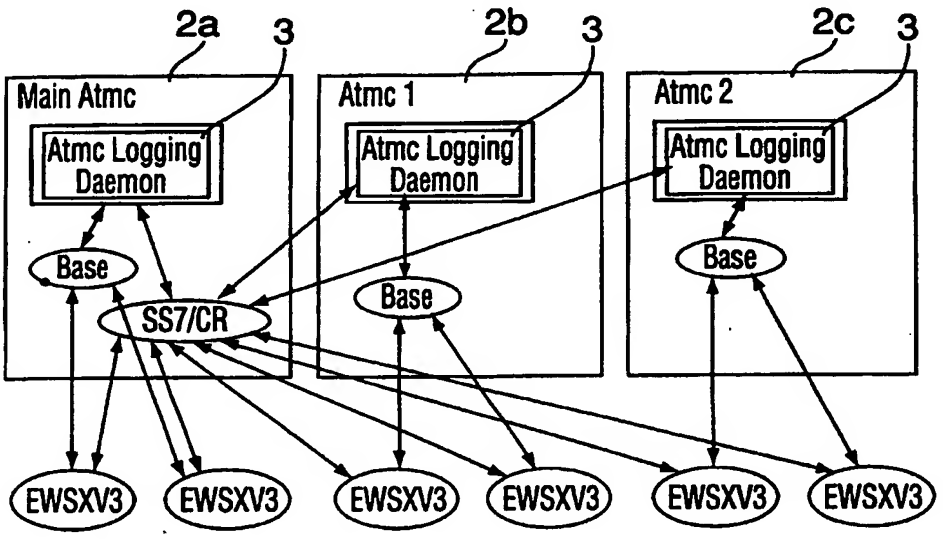
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: ZELTSER, Rafael [CA/CA]; 932 Dresden Crescent, Ottawa, Ontario K2B 5J1 (CA). BERTIN, Greg [CA/CA]; Apartment 709, 1310 Pinecrest Road, Ottawa, Ontario K2C 3W8 (CA). HAMMERSCHMIDT, George [CA/CA]; 87 Hewitt Way, Kanata, Ontario K2L 3R2 (CA).

(74) Agent: MITCHELL, Richard, J.; Marks & Clerk, P.O. Box 957, Station B, Ottawa, Ontario K1P 5S7 (CA).

(81) Designated States: CA, JP, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: TRANSACTION ROLL FORWARD



(57) Abstract

In a method of monitoring transactions in a network application, a transaction log is maintained about a node. During playback, the application will replay the original knowledge context of the transaction. If this replayed mechanism fails, a transaction rollback mechanism is invoked to restore both the application and the node to a mutually consistent state.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

TRANSACTION ROLL FORWARD

This invention relates to a method of monitoring network transactions, for example, in a wide area network controlled by network management software.

The Telecommunications Management Network (TMN) model, developed by the International Telecommunications Union (ITU), defines three areas: Element Management, Network Management and Service Management.

The Element Manager Layer controls the information exchange between network elements or groups of network elements. Functions within this layer communicate with the Network Management Layer by relaying element status and performance to a network management system (NMS).

The Network Management Layer in the TMN Model covers the end-to-end management of an entire network. It contains an overview of the network and knows how elements relate to each other. Focusing on network usage, traffic patterns, bandwidth availability and performance monitoring, the Network Management Layer interacts with the Service Management Layer by exchanging network status information for customer and service data. Responsibilities of the Network Management Layer include: controlling and coordinating the network view of all network elements within its domain, provisioning and modifying network capabilities for customer-service support, maintaining statistical, log and other data about the network and interacting with the Service Management, Layer on performance, usage and availability.

The Service Management Layer is the service provider's first point of contact with customers for all service transactions, including service requests and fault reporting, interaction with other service providers, and maintaining statistical data and interaction between services.

The present state of the art transaction roll-forward on a network element is to log all node (network element) affecting transactions from the last node backup. In the event of a failure all logs are re-applied to the node. A major shortcoming to this approach is the fact that the node has a limited context. The node logs only incorporate node affecting transaction, not necessarily element or network management transactions. For example an element management transaction "X" may be composed of node transactions

a, b and c. In the event of a failure the node will try and replay a, b and c. If a fails, b and c should not be applied because they are part of X which needs to be fully applied or not at all. The node logs have no knowledge of X. The result will be that the node and the applications supporting the node are in a inconsistent state.

The basic problem stems from the fact that a network is dynamic. A connection that was present yesterday may not exist today. It is therefore insufficient merely to the database.

An object of the invention is to overcome this problem.

According to the present invention there is provided a method of monitoring transactions in a network application, wherein a transaction log is maintained about a node, and during playback the application replays the original knowledge context of the transaction, and if this replayed mechanism fails, a transaction rollback mechanism is invoked to restore both the application and the node to a mutually consistent state.

By implementing centralized transaction logging the invention now maintains a log similar to the nodes. The invention implements a playback which relies on the application to replay the original knowledge of the context of the transaction. If this "replayed" transaction fails than the applications transaction rollback mechanism is invoked and restores both the application and the node to a mutually consistent state. When the playback is implemented, steps are taken to ensure that the log is played back in a synchronized fashion to ensure node and application consistency.

The invention can be applied to any situation where a node is modified by a multi stage transaction and may at a later time need to be restored.

The invention will now be described in more detail, by way of example only, with reference to the accompanying drawings, in which:-

Figure 1 is an overview of an ATM network to which the inventive system may be applied;

Figure 2 is an overview of a system in accordance with the invention;

Figure 3 shows a roll forward of transactions;

Figure 4 shows a cross connect creation; and

Figure 5 shows a cross connect roll forward.

Figure 1 shows the general architecture of a system to which the invention is applicable. An ATM network manager is managed by a commander 2 implementing the present invention, which communicates with the network manager via a protocol Qs, which is a published interface from Newbridge Networks Corporation.

Roll forward, shown in Figure 2, is required on a node after a catastrophe. The major assumptions behind this design are:

1. The customer has a failing switch
2. Returning as many components of the switch to service as fast as possible is by far the most important task.
3. The switch must be in a consistent state at the end of the recovery to reduce requirement for manual intervention.
4. A clear log of failed RollForward transactions must be provided to the customer.

In Figure 2, there are several commanders, 2a, 2b, 2c. The logging daemon 3 is responsible for logging the transaction, playing back the transaction, backing up the switch and restoring the transaction. The data required for logging and playback is the switch address, the time, the application, the user id, description and data.

This feature is intended for use by customers when a catastrophic event has caused the EWSX switch to require a reboot from a backup. This feature enables returning the Node to the state it was just before the catastrophe by performing a roll forward of transactions sent to the Node since the last backup.

The ATM-C Roll Forward is a product which supports the Siemens EWSXpress V3 ATM switch and applications. It provides a mechanism for the recovery of the Node from a catastrophe via the roll forward of transactions performed on the Node since a backup was performed.

A central process on the ATMC receives messages containing information on all transactions performed on the Node. The feature logs the transactions in separate file for each node. Applications, such as CR, SS7 etc. use the ATMC API Logging function to send the logging messages to the proper ATMC controlling the Node. It is assumed that

only transactions that are successfully applied to the node are logged. This implies that the logging of transactions must be done by the application only at the end of the transaction. The transaction itself contains a time stamp indicating when it has started. This information is used for the sequence of roll forward.

The Roll Forward of transactions enables the restoration of the Node Database to its state from before a catastrophic failure. Proper recovery of all the data requires synchronization between all the applications controlling the Node. This feature controls the sequence at which transaction that were applied to the node before the crash are being send to the node. This feature reads the log file generated by the logging mechanism.

The Logging and Roll Forward mechanism of the 45190 ATMC require that all applications that are acting on the Node use both of the following interfaces:

- Logging interface to log all changes to the Node
- Roll Forward Interface - To enable roll forward of all actions performed on the node.

The logging interface is part of the CKATMCAPI class and contains the following function

```
void ATMCClient :: LogOperations (ATMCAddressType Atmc, // Address of the ATMC Used for logging
        char *NodeDn, // The DN of the Node for that operation
        time_t StartTime, // The time at which the operation started.
        char *Application, // Application Name
        char *Description, // String representation of the action - used for manual roll forward
        unsigned long DataSize, // Size of the data to be logged for the event
        void *Data, // The transaction data
        long Delay = 0, // Specify amount of time in milliseconds to wait before executing next command
        int WaitCompletion = 1 // Indicates if the Roll Forward mechanism should wait for completion of this command
before sending the next one.
        );
```

Application name : Is a string used for identification of the application logging the transaction. This is displayed to the user during the Manual roll

forward and as an identifier to the roll forward mechanism of the application performing the transaction during the roll forward.

Description: A user readable string containing all the information used by the transaction. This information allows the user to screen out transactions during roll forward.

Data: This is the binary data used for roll forward. This information is send back to the application during roll forward.

DataSize: The size of the data in bytes.

WaitCompletion: See below.

During roll forward the application will send roll forward request to all the applications that logged actions.

The following pseudo code illustrates a typical applications written for roll forward:

```
Fd = RollForwardOpen ()
for ever {
        select on fd and others
        if (message on fd) { // E.G. Roll Forward
                RollForwardGetTransaction
                Process Transaction
                RollForwardSendConfirmation
        } else { // Message from other sources
                something else
    }
}
```

Following are the interfaces for roll forward:

int RollForwardOpen ()

This interface opens a socket connection that accepts messages from the Roll forward Application. The return value is a file descriptor that can be use in a select statement.

int RollForwardHasTransaction ()

This application returns 1 if there is an action waiting for roll forward in the message queue. 0 is returned if there are no messages in the queue. This is not an indication that the roll forward is done.

int RollForwardGetTransaction (unsigned short &DataSize, void * &Data, int &NeedConfirmation)

This enables the application to get the next action from the roll forward interface. The first parameter is the size of the logged data. The second parameter a void pointer to the logged Data. The third parameter indicates whether the roll forward process is going to wait for a completion message from the application, when it is set to 1 the application must use RollForwardSendConfirmation once it is safe to roll forward the next transaction. This is set to 1 only if the WaitCompletion flag was set to 1 when the application has logged the transaction.

The return value of this command is: 0 - Fail 1 - Data Available 2 - Done with roll forward

void RollForwardSendConfirmation (int Status = 1)

This command must be used if the WaitCompletion Flag was set to 1 in the RollForwardGetAction request. Use this function to indicate to the roll forward mechanism that the application is ready to receive the next transaction. This function is used only by applications that set the WaitCompletion flag in the logging record.

When the Status flag is set to 0 the RollForward mechanism assumes that the transaction has failed and will generate the appropriate log to the customer.

On startup the roll forward application will get the information of the node requiring roll forward, the time at which the last backup has been performed. (This

feature assumes that the backup generation was already applied to the node). Then, It will scan the log file and determine the list of applications that were active during the time from the backup.

The roll forward application will attempt to connect to all the applications obtained in the previous step. If any of the application fails to respond the roll forward application prompts the user to start the application manually on the appropriate machine.

As long as not all transactions from the log have been rolled forward the roll forward application will:

- send transactions to the appropriate application
- Wait for the proper reply.

During manual roll forward the RollForward application prompts the user before each transaction is send to the other applications for execution.

As long as not all actions from the log have been rolled forward the RollForward application will:

- Prompt the user for the execution of the next command
- Abort or skip command as per user input or
- send actions to the appropriate application
- Wait for the proper reply.

See Figure 2 above

The Roll Forward application will send a Finish message to all the other applications.

It is assumed that all logged operations were:

Authenticated - That during the initial invocation of the transaction the application has verified that the user have the adequate credentials that enable him to carry on the specific operation.

Successful - During the roll forward of the commands it is also assumed that all commands logged through the logging mechanism were successfully applied to the Switch.

Error recovery during roll forward is via the existing transaction roll backward of the applications. Transaction roll backward is designed as part of standard transaction handling. During roll forward the built in Roll Backward of transactions should work as during normal operation. Since roll forward is done for transactions that were successfully applied to the node, it is anticipated that this mechanism will rarely be invoked. However, in the case of hardware failure, during roll forward, of one of the modules on the Node a transaction may abort. In this case the transaction should behave the same way as it would under normal conditions - that is roll backward. This will leave the Switch in the most consistent way possible. Logs indicating the failure will appear in a standard log. Since we do not necessarily have a GUI connected to the applications the SendConfirmation message to the RollForward mechanism should indicate the failure so that the Roll Forward mechanism can generate additional required logs.

It is known that this mechanism will not manage to completely restore the Node to its original condition under the following conditions:

1. A race condition between two applications or transactions while setting an attribute for the same object. In this case there is no way to determine which application will be the last to set the attribute, and therefor to assure proper operation.
2. There is a physical problem with one of the components on the Switch.

The following applications are known to be using the logging and roll forward API stubs at the release 2 time frame:

- ATMC base.
- Call Routing
- SS7
- Q3PS

- Autocraft

The following example is an example of creation of a cross connect. This type of operation is a very typical one on the ATMC and at any given time we might have multiple cross connect transactions in operation.

As can be seen from the above figures, using cross Q3 roll forward will result in reversing the state of the two object. That is the connection represented by Object 1 will be inactive while the connection represented by Object 2 is going to be active. This is in contrast to the original state of the connections where the connection represented by Object 1 was active and the connection represented by object 2 was inactive.

Q3 logging at the object level is the original proposal prepared for CR/SS7. In this design the application is logging the Q3 objects just before they send the command to the Node. Using Vertel this implies that before sending commands to the node the application can translate the Vertel object to a BER representation using the Vertel "Ber_Encode" template then send the information to the logging mechanism using a function as follows:

```
void ATMCClient :: LogOperations (ATMCAddressType Atmc, // Address of the ATMC Used for logging
        char *NodeDn, // The DN of the Node for that operation
        time_t StartTime, // The time at which the operation started.
        char *Application, // Application Name
        char *Description, // String representation of the action - used for manual roll forward
        long DataSize, // Size of the data to be logged for the event
        void *Data, // The BER of the transaction object
        long Delay = 0, // Specify amount of time in milliseconds to wait before executing next command
        int WaitCompletion = 1 // Indicates if the Roll Forward mechanism should wait for completion of this command
before sending the next one.
        );
```

During the roll forward of the transaction the Roll Forward agent will perform similar tasks to the ones described above. The only difference between is that instead of sending the Q3 actions to the applications for execution the BER data of the Q3 actions is converted to a Vertel Object representing the Q3 command (Using features of the Vertel Stack T.B.D) and the action is than sent to the Node for execution. Q3 is a standards

based interface for managing network elements. The Siemens EWSX V3 is managed via a Q3 interface.

Logging at the Socket level is similar to the design described in the previous section, except that instead of the application calling a special function for the purpose of logging of the commands the data is intercepted and logged at the MP120 Daemon. This design has the following problems:

- Requires modification of the MP120 (Vertel).
- Does not support any other stack such as the one used by Q3PS.
- Selection of 'actions' during the roll forward is very hard since there is no information as for the generating application and user (Official requirement).

The invention is applicable to any situation where a node is modified by a multi stage transaction emitted from multiple applications and may at a later time need to be restored.

Claims:

1. A method of monitoring transactions in a network application, wherein a transaction log is maintained about a node, and during playback the application replays the original knowledge context of the transaction, and if this replayed mechanism fails, a transaction rollback mechanism is invoked to restore both the application and the node to a mutually consistent state.

2. A method as claimed in claim 1, wherein a centralized log is maintained similar to the log at the nodes, and when a playback is implemented the log is played back in a synchronized manner to ensure node and application consistency.

3. A method as claimed in claim 2, wherein said centralized log is maintained on a network element manager.

4. A method as claimed in claim 3, wherein said log stores the switch address, time, application, user id, description, and data.

5. A method as claimed in claim 4, wherein said transaction log is maintained with the aid of a logging daemon.

6. A method as claimed in claim 1, wherein said mutually consistent state is the state of the network immediately prior to a catastrophic failure.

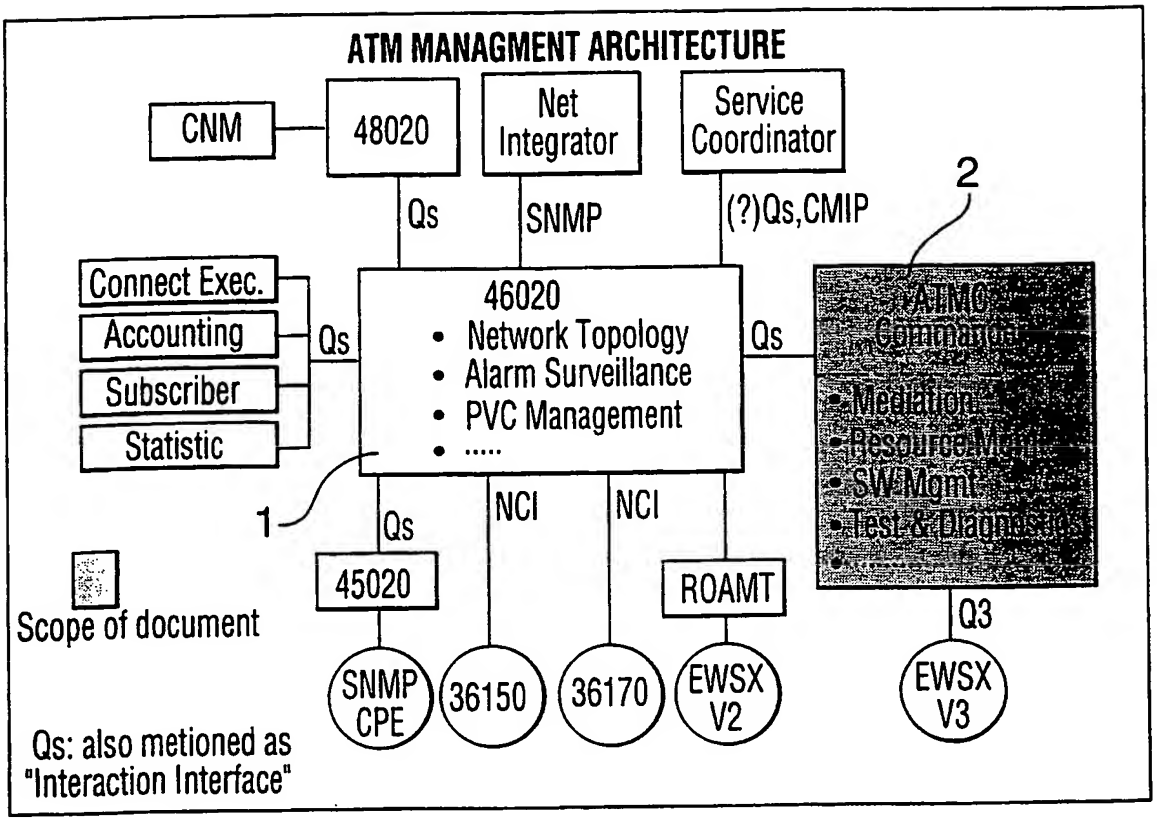

ATM MANAGMENT ARCHITECTURE
CNM
48020
Net Integrator
Service Coordinator
Qs
SNMP
(?)Qs,CMIP
2
Connect Exec.
Accounting
Subscriber
Statistic
Qs
46020
• Network Topology
• Alarm Surveillance
• PVC Management
• .....
Qs
ATMC
Command
• Mediation
• Resource Mgmt
• SW Mgmt
• Test & Diagnosis
1
Qs
NCI
NCI
45020
ROAMT
Q3
Scope of document
SNMP CPE
36150
36170
EWSX V2
EWSX V3
Qs: also metioned as "Interaction Interface"

FIG. 1

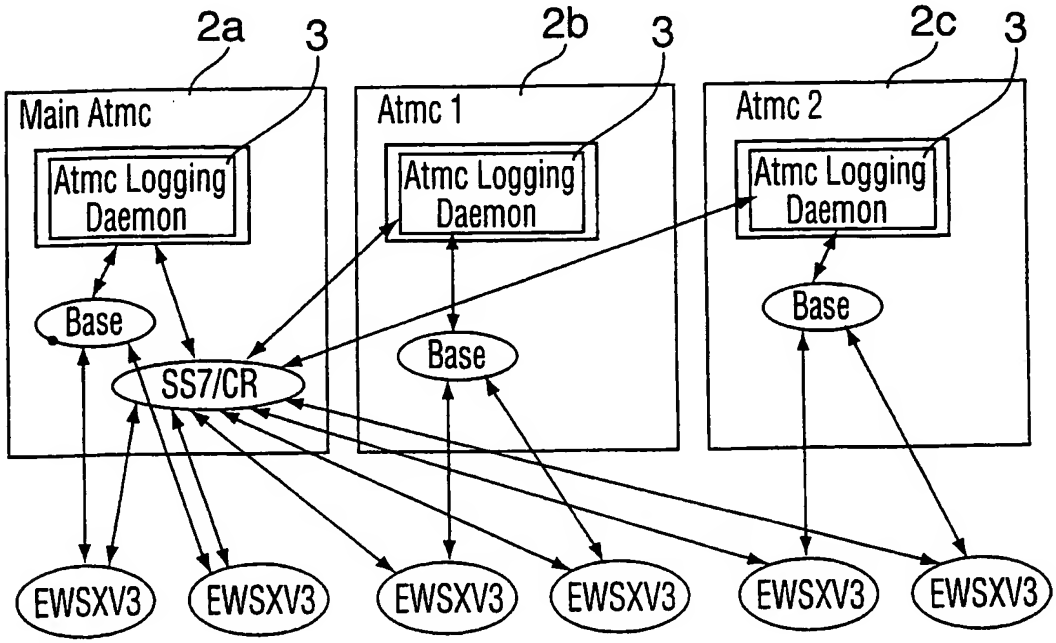

2a
3
2b
3
2c
3
Main Atmc
Atmc 1
Atmc 2
Atmc Logging Daemon
Atmc Logging Daemon
Atmc Logging Daemon
Base
Base
Base
SS7/CR
EWSXV3
EWSXV3
EWSXV3
EWSXV3
EWSXV3
EWSXV3

FIG. 2

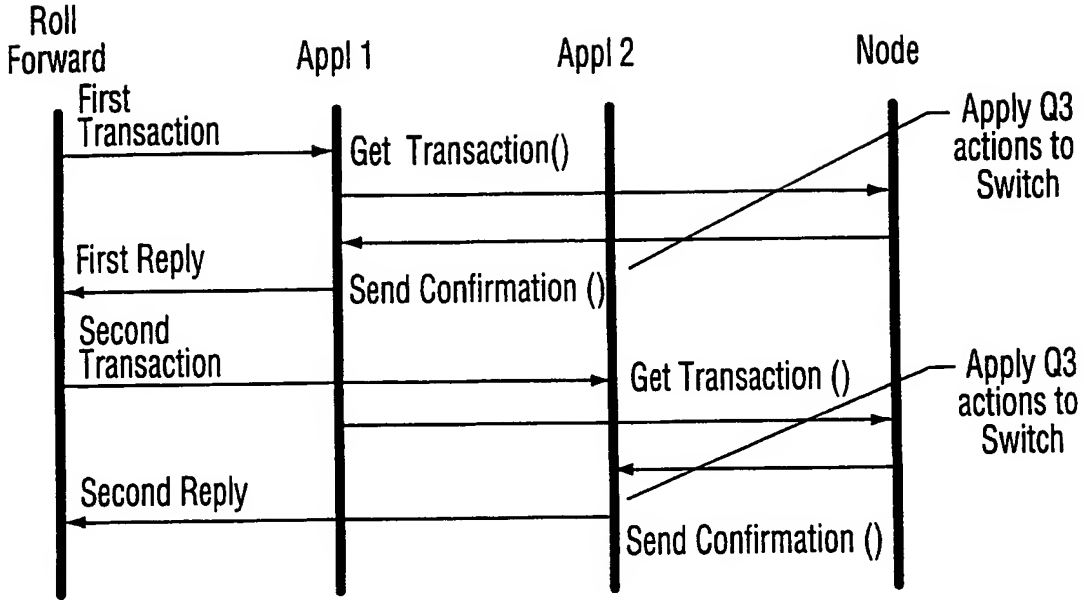

Roll Forward
Appl 1
Appl 2
Node
First Transaction
Get Transaction()
Apply Q3 actions to Switch
First Reply
Send Confirmation ()
Second Transaction
Get Transaction ()
Apply Q3 actions to Switch
Second Reply
Send Confirmation ()

FIG. 3

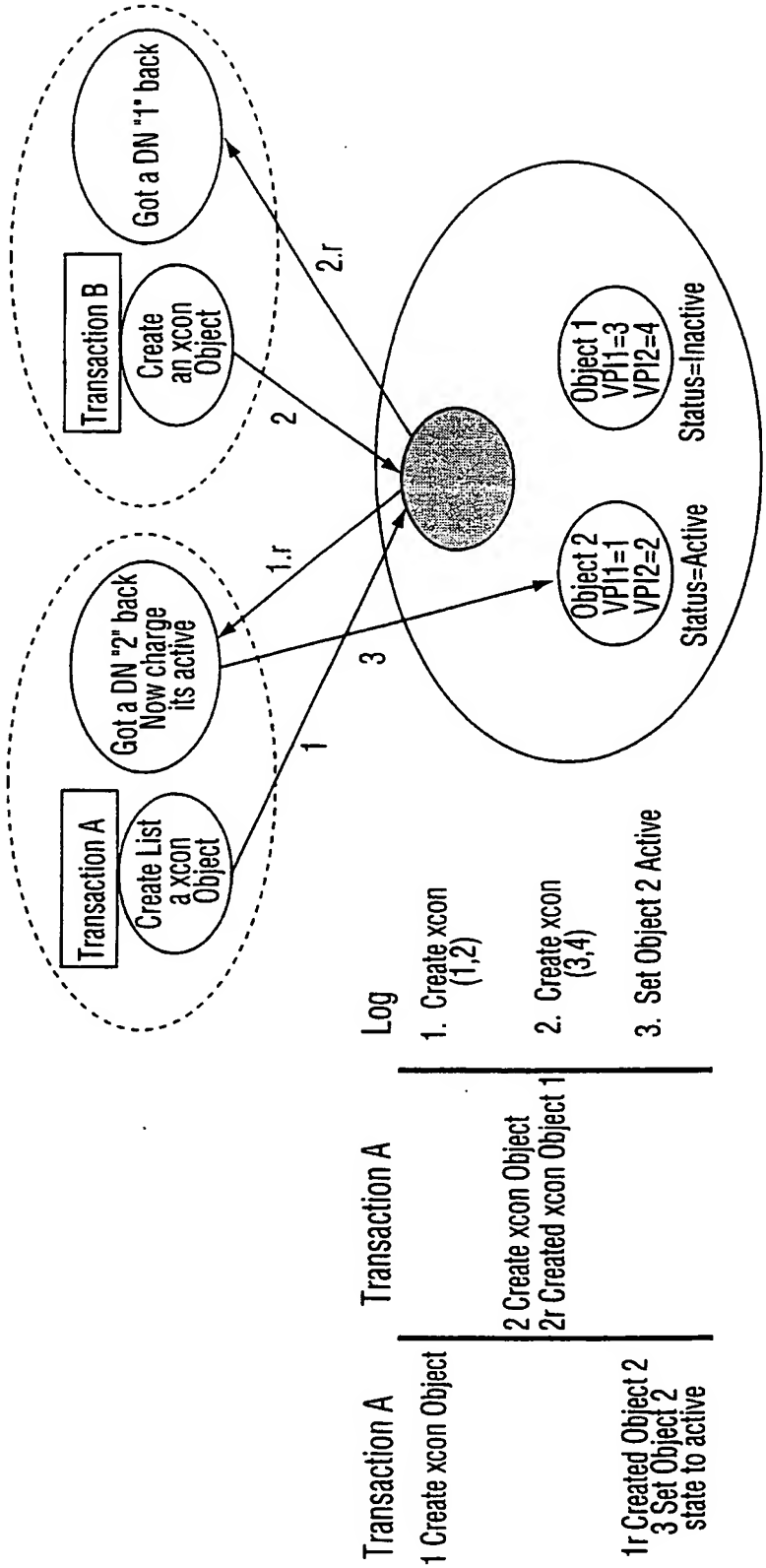

Transaction A
Create List a xcon Object
Got a DN "2" back Now charge its active
Transaction B
Create an xcon Object
Got a DN "1" back
1
1.r
2
2.r
3
Object 2 VPI1=1 VPI2=2
Status=Active
Object 1 VPI1=3 VPI2=4
Status=Inactive
Log
1. Create xcon (1,2)
2. Create xcon (3,4)
3. Set Object 2 Active
Transaction A
1 Create xcon Object
Transaction A
2 Create xcon Object
2r Created xcon Object 1
1r Created Object 2
3 Set Object 2 state to active

FIG. 4

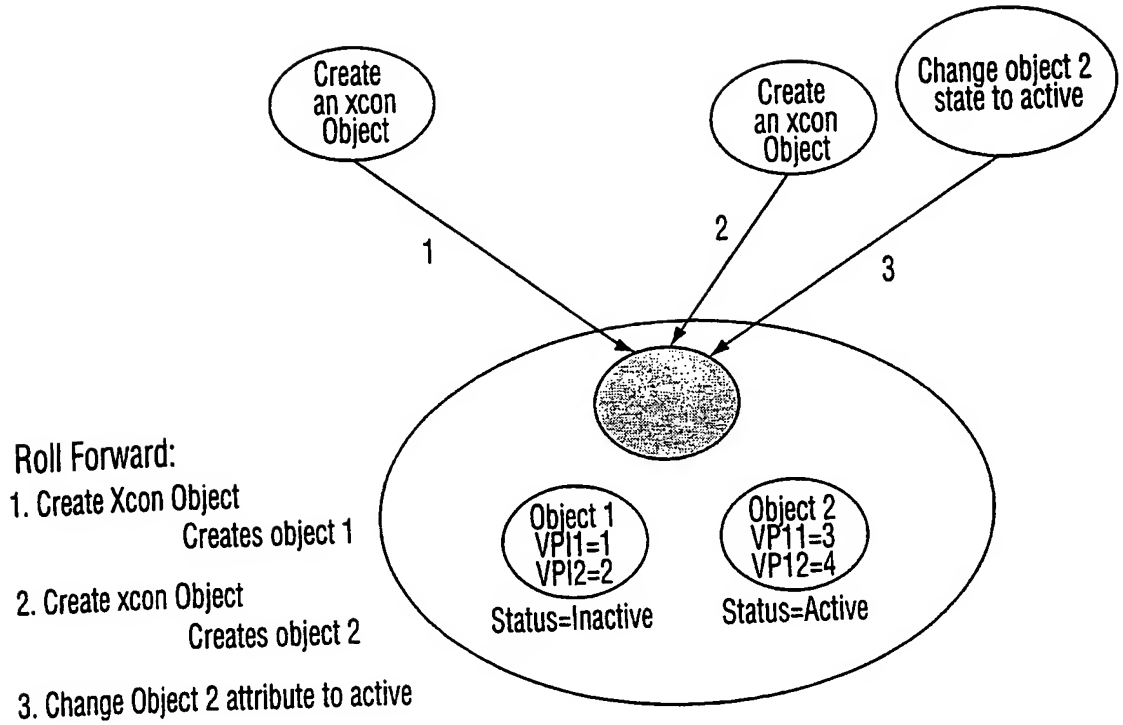

Create an xcon Object
Create an xcon Object
Change object 2 state to active
1
2
3
Roll Forward:
1. Create Xcon Object
Creates object 1
2. Create xcon Object
Creates object 2
3. Change Object 2 attribute to active
Object 1
VPI1=1
VPI2=2
Status=Inactive
Object 2
VP11=3
VP12=4
Status=Active

FIG. 5

INTERNATIONAL SEARCH REPORT

Internati Application No
PCT/CA 98/01074

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 H04Q11/04 H04L29/06 H04Q3/00 G06F11/14

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 H04Q H04L G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	RAVI PRAKASH ET AL: "LOW-COST CHECKPOINTING AND FAILURE RECOVERY IN MOBILE COMPUTING SYSTEMS" IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, vol. 7, no. 10, October 1996, pages 1035-1048, XP000639593 see paragraph 1 ---	1-6
A	AGARWAL A ET AL: "A UNIFIED APPROACH TO FAULT-TOLERANCE IN COMMUNICATION PROTOCOLS BASED ON RECOVERY PROCEDURES" IEEE / ACM TRANSACTIONS ON NETWORKING, vol. 4, no. 5, October 1996, pages 785-795, XP000631091 see abstract --- -/--	1-6

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
9 March 1999	19/03/1999

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Staessen, B

1

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Internat' Application No
PCT/CA 98/01074

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	YOSHIAKI KAKUDA: "A RECOVERY SEQUENCE GENERATION SYSTEM FOR DESIGN OF RECOVERABLE PROTOCOLS" IEICE TRANSACTIONS, vol. E74, no. 6, 1 June 1991, pages 1715-1726, XP000262327 see abstract	1-6
A	PATENT ABSTRACTS OF JAPAN vol. 098, no. 002, 30 January 1998 & JP 09 259098 A (TOSHIBA CORP), 3 October 1997 see abstract	4,6
A	MUSTAQUE AHAMAD ET AL: "USING CHECKPOINTS TO LOCALIZE THE EFFECTS OF FAULTS IN DISTRIBUTED SYSTEMS" PROCEEDINGS OF THE SYMPOSIUM ON RELIABLE DISTRIBUTED SYSTEMS, SEATTLE, 10 - 12 OCT., 1989, no. SYMP. 8, 10 October 1989, pages 2-11, XP000088982 INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS	2
A	PARIS J -F: "A HIGHLY AVAILABLE REPLICATION CONTROL PROTOCOL USING VOLATILE WITNESSES" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON DISTRIBUTED COMPUTIN SYSTEMS, POZNAN, POLAND, JUNE 21 - 24, 1994, no. CONF. 14, 21 June 1994, pages 536-543, XP000489116 INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS see abstract	1

1